Empirical Covering Number Analysis and Symmetrization

Mathematical Analysis of Machine Learning Algorithms
(Chapter 4)

Metric Covering Number

We introduce metric covering numbers on a general pseudometrics space as follows.

Definition 1 (Def 4.1)

Let $(\mathcal{V}, d)$ be a pseudometric space with metric $d(\cdot, \cdot)$. A finite set $\mathcal{G}(\epsilon) \subset \mathcal{V}$ is an ϵ cover (or ϵ net) of $\mathcal{G} \subset \mathcal{V}$ if, for all $\phi \in \mathcal{G}$, there exists $\phi' \in \mathcal{G}(\epsilon)$ so that $d(\phi', \phi) \leq \epsilon$. The ϵ-covering number of $\mathcal{G}$ with metric d is the smallest cardinality $N(\epsilon, \mathcal{G}, d)$ of such $\mathcal{G}(\epsilon)$. The number $\ln N(\epsilon, \mathcal{G}, d)$ is called the ϵ-entropy.

For a function class $\mathcal{G}$ with seminorm $L_p(\mathcal{D})$ ($p \geq 1$)

$$\|f - f'\|_{L_p(\mathcal{D})} = \left[\mathbb{E}_{Z \sim \mathcal{D}} |f(Z) - f'(Z)|^p\right]^{1/p},$$

the corresponding $L_p(\mathcal{D})$-covering number is $N(\epsilon, \mathcal{G}, L_p(\mathcal{D}))$.

Relation to Bracketing Number

The L_∞ bracketing cover is equivalent to L_∞ cover. Therefore uniform convergence results in Chapter 3 holds for L_∞ cover.

Proposition 2 (Prop 4.3)

$$N_{LB}(\epsilon, \mathcal{G}, L_1(\mathcal{D})) \leq N_{[]}(\epsilon, \mathcal{G}, L_\infty(\mathcal{D})) = N(\epsilon/2, \mathcal{G}, L_\infty(\mathcal{D})).$$

However, one cannot derive uniform convergence result based on L_p covering number with $p < \infty$, although one can derive such results with bracketing number (see Chapter 3).

We need to work with empirical/uniform L_p covering number to obtain uniform convergence results.

Empirical and Uniform Covering Number

Definition 3 (Def 4.4)

Given an empirical distribution $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$, we define the pseudometric $d = L_p(\mathcal{S}_n)$ as

$$d(\phi, \phi') = \left[\frac{1}{n} \sum_{i=1}^{n} |\phi(Z_i) - \phi'(Z_i)|^p \right]^{1/p}.$$

The corresponding metric covering number $N(\epsilon, \mathcal{G}, L_p(\mathcal{S}_n))$ is referred to as the empirical L_p covering number. Given n, the largest L_p covering number over empirical distribution $\mathcal{S}_n$ is referred to as the uniform L_p covering number

$$N_p(\epsilon, \mathcal{G}, n) = \sup_{\mathcal{S}_n} N(\epsilon, \mathcal{G}, L_p(\mathcal{S}_n)).$$

Properties

Proposition 4

For $1 \leq p \leq q$, *we have*

$$\begin{aligned} N(\epsilon, \mathcal{G}, L_p(\mathcal{S}_n)) &\leq N(\epsilon, \mathcal{G}, L_q(\mathcal{S}_n)), \\ N_p(\epsilon, \mathcal{G}, n) &\leq N_q(\epsilon, \mathcal{G}, n). \end{aligned}$$

We will later show that the uniform L_1 covering number can be used to obtain uniform convergence and oracle inequalities.

Example

Example 5

Consider $\{0, 1\}$ valued linear classifiers in d dimension of the form $f(w, x) = \mathbb{1}(w^\top x \geq 0)$, where $w \in \Omega = \mathbb{R}^d$ and $\in \mathcal{X} = \mathbb{R}^d$. Let $Y \in \{0, 1\}$, then classification error is $\phi(w, z) = \mathbb{1}(f(w, x) \neq y)$, where $z = (x, y)$.

- Difficult to obtain bracketing number.
- However it is easy to obtain L_∞ empirical covering number:

$$N_\infty(\mathcal{G}, \epsilon = 0, n) \leq (2n)^d.$$

Symmetrization: Notations

Let $Z = (X, Y)$. Consider

- Training data $\mathcal{S}_n = \{Z_1, \ldots, Z_n\}$, drawn independently from $\mathcal{D}$
- Validation data $\mathcal{S}'_n = \{Z'_1, \ldots, Z'_n\}$, drawn independently from $\mathcal{D}$.

Given a function $f(Z)$, define the training loss and the validation loss:

$$f(\mathcal{S}_n) = \frac{1}{n} \sum_{Z \in \mathcal{S}_n} f(Z), \quad f(\mathcal{S}'_n) = \frac{1}{n} \sum_{Z \in \mathcal{S}'_n} f(Z).$$

Let $\mathcal{F}$ be a function class. Consider n iid Bernoulli random variables $\sigma_i \in \{\pm 1\}$, where $\Pr(\sigma_i = 1) = \Pr(\sigma_i = -1) = 0.5$. The symmetrized empirical process is

$$f(\sigma, \mathcal{S}_n) = \frac{1}{n} \sum_{i=1}^{n} \sigma_i f(Z_i) \qquad f \in \mathcal{F},$$

where the randomness is with respect to $\mathcal{S}_n = \{Z_i\}$ and $\sigma = \{\sigma_i\}$.

Symmetrization Lemma

Lemma 6 (Simplified from Lem 4.8)

Consider a real valued function family $\mathcal{F} = \{f : \mathcal{Z} \to \mathbb{R}\}$. Assume there exists a function $\epsilon_n : (0,1) \to \mathbb{R}$ so that with probability at least $1-\delta$:

$$\forall f \in \mathcal{F}, \quad f(\sigma, \mathcal{S}_n) \leq \epsilon_n(\delta),$$

where the randomness is over both $\mathcal{S}_n \sim \mathcal{D}^n$ and σ. Then with probability at least $1-\delta$ over independent random data $(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}$:

$$\forall f \in \mathcal{F}, \quad f(\mathcal{S}'_n) \leq f(\mathcal{S}_n) + 2\epsilon_n(\delta/2).$$

Convergence of training error to validation error can be obtained from uniform convergence of symmetrized empirical process.

Proof of Lemma 7

Consider independent random samples $(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}$. The distribution of $f(\mathcal{S}_n) - f(\mathcal{S}'_n)$ is the same as that of $f(\sigma, \mathcal{S}_n) - f(\sigma, \mathcal{S}'_n)$, and the latter contains additional randomness from Bernoulli random variables σ, drawn independently of $(\mathcal{S}_n, \mathcal{S}'_n)$. It follows that

$$\begin{aligned}
& \Pr\left(\exists f \in \mathcal{F}, f(\mathcal{S}'_n) > f(\mathcal{S}_n) + 2\epsilon_n(\delta/2)\right) \\
= & \Pr\left(\exists f \in \mathcal{F}, f(\sigma, \mathcal{S}'_n) > f(\sigma, \mathcal{S}_n) + 2\epsilon_n(\delta/2)\right) \\
\leq & 2\Pr\left(\exists f \in \mathcal{F}, f(\sigma, \mathcal{S}_n) > \psi(f, \mathcal{S}_n) + \epsilon_n(\delta/2)\right) \leq 2(\delta/2).
\end{aligned}$$

In the above derivation, the first equation used the fact that $f(\mathcal{S}_n) - f(\mathcal{S}'_n)$ and $f(\sigma, \mathcal{S}_n) - f(\sigma, \mathcal{S}'_n)$ have the same distributions. The first inequality used the union bound, and the symmetry of $-f(\sigma, \mathcal{S}_n)$ and $f(\sigma, \mathcal{S}_n)$.

From Validation Loss to Test Loss

Lemma 7 (Simplification of Lem 4.11)

Assume that Lemma 7 holds. With probability at least $1-\delta_1$ *over independent random data* $(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}$*:*

$$\forall f \in \mathcal{F}, \quad f(\mathcal{S}'_n) \leq f(\mathcal{S}_n) + 2\epsilon_n(\delta_1/2).$$

Moreover, assume $\forall f \in \mathcal{F}$*, we have with probability* $1-\delta_2$ *over randomly drawn* $\mathcal{S}'_n \sim \mathcal{D}$*:*

$$f(\mathcal{D}) \leq f(\mathcal{S}'_n) + \epsilon'_n(\delta_2), \tag{1}$$

where $f(\mathcal{D}) = \mathbb{E}_{Z\sim\mathcal{D}} f(Z)$*. Then the following uniform convergence statement holds. With probability at least* $1-\delta_1-\delta_2$*,*

$$\forall f \in \mathcal{F}: \; f(\mathcal{D}) \leq f(\mathcal{S}_n) + 2\epsilon_n(\delta_1/2) + \epsilon'_n(\delta_2).$$

Note that we do not need uniform convergence in (1).

Proof of Lemma 8 (I/II)

Let $Q(f, \mathcal{S}_n) = f(\mathcal{D}) - f(\mathcal{S}_n) - (2\epsilon_n(\delta_1/2) + \epsilon'_n(\delta_2))$, and let E be the event that $\sup_{f \in \mathcal{F}} Q(f, \mathcal{S}_n) \leq 0$. We pick $\hat{f}_{\mathcal{S}_n} \in \mathcal{F}$ so that

- If E holds, choose $\hat{f}_{\mathcal{S}_n}$ so that $Q(\hat{f}_{\mathcal{S}_n}, \mathcal{S}_n) \leq 0$.
- If E does not hold, choose $\hat{f}_{\mathcal{S}_n}$ so that $Q(\hat{f}_{\mathcal{S}_n}, \mathcal{S}_n) > 0$.

Consider sample $(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}$. The uniform convergence condition implies that with probability at least $1 - \delta_1$, the following event holds:

$$E_1 : \quad \hat{f}_{\mathcal{S}_n}(\mathcal{S}'_n) \leq \hat{f}_{\mathcal{S}_n}(\mathcal{S}_n) + 2\epsilon_n(\delta_1/2).$$

The condition of the theorem also implies that with probability at least $1 - \delta_2$, the following event holds:

$$E_2 : \quad \hat{f}_{\mathcal{S}_n}(\mathcal{D}) \leq \hat{f}_{\mathcal{S}_n}(\mathcal{S}'_n) + \epsilon'_n(\delta_2).$$

Proof of Lemma 8 (II/II)

If both events E_1 and E_2 hold, then

$$\hat{f}_{\mathcal{S}_n}(\mathcal{D}) \leq \hat{f}_{\mathcal{S}_n}(\mathcal{S}'_n) + \epsilon'_n(\delta_2) \leq \hat{f}_{\mathcal{S}_n}(\mathcal{S}_n) + 2\epsilon_n(\delta_1/2) + \epsilon'_n(\delta_2).$$

From the definition of $\hat{f}_{\mathcal{S}_n}$, we know that E holds.
Therefore

$$\Pr(E) \geq \Pr(E_1 \& E_2) \geq 1 - \delta_1 - \delta_2.$$

Uniform Convergence with Uniform L_1 Covering

Using the same notations of Chapter 3, we consider a function class

$$\mathcal{G} = \{\phi(w, z) : w \in \Omega\}.$$

Theorem 8 (Additive Bound in Thm 4.12)

Assume that $\phi(w, z) \in [0, 1]$ *for all* w *and* z*. Then given* $\delta \in (0, 1)$*, with probability at least* $1 - \delta$*, the following inequality holds:*

$$\forall w \in \Omega : \phi(w, \mathcal{D}) \leq \phi(w, \mathcal{S}_n) + \epsilon_n(\delta),$$

where

$$\epsilon_n(\delta) = \inf_{\epsilon>0} \left[2\epsilon + 3\sqrt{\frac{\ln(3N_1(\epsilon, \mathcal{G}, 2n)/\delta)}{2n}} \right].$$

Proof of Theorem 9 (I/II)

Let $\mathcal{F} = \{f(z) = \phi(w, z) - 0.5 : w \in \Omega\}$. Given $\mathcal{S}_n$, we consider an ϵ-$L_1(\mathcal{S}_n)$ cover $\mathcal{F}_\epsilon(\mathcal{S}_n)$ of $\mathcal{F}$, of size no more $N = N_1(\epsilon, \mathcal{G}, n)$. We may assume that $f(Z_i) \in [-0.5, 0.5]$ for $f \in \mathcal{F}_\epsilon(\mathcal{S}_n)$. From Corollary 2.27 (with $a_i = 0.5$) and the union bound, we obtain the following uniform convergence result over $\mathcal{F}_\epsilon(\mathcal{S}_n)$. With probability $1 - \delta$:

$$\forall f \in \mathcal{F}_\epsilon(\mathcal{S}_n) : f(\sigma, \mathcal{S}_n) \leq \sqrt{\frac{\ln(N/\delta)}{2n}}.$$

Since for all $f \in \mathcal{F}$, we can find $f' \in \mathcal{F}_\epsilon(\mathcal{S}_n)$ so that $n^{-1} \sum_{Z \in \mathcal{S}_n} |f(Z) - f'(Z)| \leq \epsilon$ for all $Z \in \mathcal{S}_n$. It follows that

$$f(\sigma, \mathcal{S}_n) \leq f'(\sigma, \mathcal{S}_n) + \epsilon \leq \epsilon + \sqrt{\frac{\ln(N/\delta)}{2n}}.$$

Proof of Theorem 9 (II/II)

Using Lemma 7 with $\psi = 0$, this uniform convergence result for the symmetrized empirical process implies the following uniform convergence result. With probability at least $1 - \delta_1$ over $(\mathcal{S}_n, \mathcal{S}'_n) \sim \mathcal{D}^{2n}$:

$$\forall w \in \Omega : \underbrace{\phi(w, \mathcal{S}'_n)}_{\text{validation error}} \leq \underbrace{\phi(w, \mathcal{S}_n)}_{\text{training error}} + \underbrace{2\epsilon + \sqrt{\frac{2\ln(2N/\delta_1)}{n}}}_{2\epsilon_n(\delta_1/2)} .$$

The standard additive Chernoff bound implies that for all $w \in \Omega$, with probability at least $1 - \delta_2$:

$$\underbrace{\phi(w, \mathcal{D})}_{\text{test error}} \leq \underbrace{\phi(w, \mathcal{S}'_n)}_{\text{validation error}} + \underbrace{\sqrt{\frac{\ln(1/\delta_2)}{2n}}}_{\epsilon'_n(\delta_2)} .$$

Therefore in Lemma 8, we can take symbols as defined above, together with $\delta_1 = 2\delta/3$ and $\delta_2 = \delta/3$ to obtain the desired bound.

Oracle Inequality

Corollary 9 (Additive Bound in Cor 4.13)

If $\phi(w, z) \in [0, 1]$. *Let* $\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$. *With probability at least* $1 - \delta$, *the approximate ERM method* (3.3) *satisfies the (additive) oracle inequality:*

$$\begin{aligned} \mathbb{E}_{Z\sim\mathcal{D}}\phi(\hat{w}, Z) \leq & \inf_{w\in\Omega} \mathbb{E}_{Z\sim\mathcal{D}}\phi(w, Z) + \epsilon' \\ & + \inf_{\epsilon>0}\left[2\epsilon + \sqrt{\frac{8\ln(4N_1(\epsilon, \mathcal{G}, n)/\delta)}{n}}\right]. \end{aligned}$$

Example

Example 10 (Additive Bound, Expl 4.14)

Consider the linear classifier example in Example 6. Since

$$\ln N_\infty(\epsilon, \mathcal{G}, n) \leq d \ln(2n),$$

it follows that for the ERM method, we have the following oracle inequalities. With probability at least $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} \mathbb{1}(f(\hat{w}, X) \neq Y) \leq \inf_{w \in \mathbb{R}^d} \mathbb{E}_{\mathcal{D}} \mathbb{1}(f(w, X) \neq Y) + \sqrt{\frac{8(\ln(4/\delta) + d \ln(2n))}{n}}.$$

Example (cont)

Example 11 (Multiplicative Bound, Expl 4.14)

With probability at least $1 - \delta$:

$$\mathbb{E}_{\mathcal{D}} \mathbb{1}(f(\hat{w}, X) \neq Y) \leq \mathrm{err}_* + C \left[\sqrt{\mathrm{err}_* \frac{\ln(\delta^{-1}) + d \ln(n)}{n}} + \frac{\ln(\delta^{-1}) + d \ln(n)}{n} \right],$$

where C is an absolute constant and

$$\mathrm{err}_* = \inf_{w \in \Omega} \mathbb{E}_{\mathcal{D}} \mathbb{1}(f(w, X) \neq Y).$$

The multiplicative bound is better than the additive bound when $\mathrm{err}_* \approx 0$. Details of the derivation can be found in the book.

Vapnik-Chervonenkis Dimension

Let $\mathcal{G} = \{\phi(w, z) : w \in \Omega\}$ be a $\{0, 1\}$ valued binary function class of $z \in \mathcal{Z}$ indexed by $w \in \Omega$.

Definition 12 (VC-dimension)

We say that $\mathcal{G}$ shatters $\mathcal{S}_n$ if the number of elements $|\mathcal{G}(\mathcal{S}_n)|$ is 2^n. That is, we can always find $w \in \Omega$ so that $\phi(w, z)$ matches any arbitrary possible choice of $\{0, 1\}^n$ values at the n points. The maximum n such that $\mathcal{G}$ shatters at least one instance of $\mathcal{S}_n \in \mathcal{Z}^n$, denoted by $\text{VC}(G)$, is called the VC-dimension of $\mathcal{G}$.

Sauer's Lemma

Lemma 13 (Sauer's Lemma, Lem 4.16)

If $\text{VC}(\mathcal{G}) = d$*, then we have for all* $n > 0$ *and empirical samples* $\mathcal{S}_n = \{Z_1, \ldots, Z_n\} \in \mathcal{Z}^n$*:*

$$|\mathcal{G}(\mathcal{S}_n)| \leq \sum_{\ell=0}^{d} \binom{n}{\ell} \leq \max(2, en/d)^d.$$

Proof of Lemma 14 (I/III)

First, we prove the statement under the assumption that $|\mathcal{G}(\mathcal{S}_n)|$ is upper bounded by the number of subsets of $\mathcal{S}_n$ (including the empty set) that are shattered by $\mathcal{G}$.
Under this assumption, since any subset shattered by $\mathcal{G}$ cannot be larger than d by the definition of VC-dimension, and the number of subsets of size ℓ is $\binom{n}{\ell}$, we know that the number of subsets shattered by $\mathcal{G}$ cannot be more than $\sum_{\ell=1}^{d} \binom{n}{\ell}$.

When $n \geq d$, we have (see Exercise 4.1)

$$\sum_{\ell=0}^{d} \binom{n}{\ell} \leq (en/d)^d. \tag{2}$$

When $n \leq d$, we have $\sum_{\ell=0}^{d} \binom{n}{\ell} \leq 2^d$. This implies the desired result.

Proof of Lemma 14 (II/III)

In the following, we only need to prove the statement that $|\mathcal{G}(\mathcal{S}_n)|$ is upper bounded by the number of subsets of $\mathcal{S}_n$ that are shattered by $\mathcal{G}$. This can be proved by induction on n. When $n = 1$, one can check that the claim holds trivially.

Now assume that the claim holds for all empirical samples of size no more than $n - 1$. Consider n samples $\{Z_1, \ldots, Z_n\}$. We define

$$\begin{aligned} \phi(w, \mathcal{S}_k) =& [\phi(w, Z_1), \ldots, \phi(w, Z_k)], \\ \mathcal{G}_{n-1}(\mathcal{S}_n) =& \{[\phi(w, \mathcal{S}_{n-1}), 1] : [\phi(w, S_{n-1}), 0], [\phi(w, \mathcal{S}_{n-1}, 1] \in \mathcal{G}(\mathcal{S}_n)\} . \end{aligned}$$

Using the induction hypothesis, we know that $|\mathcal{G}_{n-1}(\mathcal{S}_n)|$ is bounded by the number of shattered subset $\mathcal{S} \subset \mathcal{S}_{n-1}$; for each shattered $\mathcal{S} \subset \mathcal{S}_{n-1}$, $\mathcal{S} \cup \{Z_n\}$ is shattered by $\mathcal{G}(\mathcal{S}_n)$ because both $[\phi(w, \mathcal{S}_{n-1}), 1]$ and $[\phi(w, S_{n-1}), 0]$ belong to $\mathcal{G}(\mathcal{S}_n)$.
Therefore $|\mathcal{G}_{n-1}(\mathcal{S}_n)|$ is no more than the number of shattered subsets of $\mathcal{S}_n$ that contains Z_n.

Proof of Lemma 14 (III/III)

Moreover, since for $\phi(w, \cdot) \in \mathcal{G}(\mathcal{S}_n) - \mathcal{G}_{n-1}(\mathcal{S}_n)$, $\phi(w, Z_n)$ is uniquely determined by its values at $\mathcal{S}_{n-1}$[1], it follows that $|\mathcal{G}(\mathcal{S}_n) - \mathcal{G}_{n-1}(\mathcal{S}_n)|$ is no more than $|\mathcal{G}(S_{n-1})|$.
By induction hypothesis, $|\mathcal{G}(\mathcal{S}_{n-1})|$ is no more than the number of shattered subsets of $\mathcal{S}_n$ that does not contain Z_n.

By combining the above two facts, $|\mathcal{G}(\mathcal{S}_n)|$ is no more than the number of shattered subsets of $\mathcal{S}_n$.

[1] If not, then both $[\phi(w, \mathcal{S}_{n-1}), 0]$ and $[\phi(w, \mathcal{S}_{n-1}), 1]$ can be achieved in $\mathcal{G}(\mathcal{S}_n) - \mathcal{G}_{n-1}(\mathcal{S}_n)$, which is impossible because by definition, we should have put $[\phi(w, \mathcal{S}_{n-1}), 1]$ in $\mathcal{G}_{n-1}(\mathcal{S}_n)$

Example of Finite VC Dimension

Proposition 14 (Prop 4.18)

Consider d-dimensional $\{0, 1\}$ *valued linear classifiers of the form*

$$\mathcal{F} = \{f_w(x) = \mathbb{1}(w^\top x \geq 0), w \in \mathbb{R}^d\},$$

we have $\mathrm{VC}(\mathcal{F}) = d$*.*
This implies that d-dimensional linear classifier

$$\mathcal{G} = \{\mathbb{1}(f_w(X) \neq Y), w \in \mathbb{R}^d\}$$

has VC dimension $\mathrm{VC}(\mathcal{G}) = d$*.*

Proof of Proposition 15

Consider $d+1$ points $x_1, \dots, x_{d+1}$. There exists $d+1$ real valued coefficients $[a_1, \dots, a_{d+1}] \neq 0$ such that $\exists a_j > 0$ and

$$a_1 x_1 + \cdots + a_{d+1} x_{d+1} = 0. \tag{3}$$

In order to show that $x_1, \dots, x_{d+1}$ cannot be shattered, we only need to show that there is no $w \in \mathbb{R}^d$ such that

$$\mathbb{1}(w^\top x_i \geq 0) = 0 \quad (a_i > 0); \qquad \mathbb{1}(w^\top x_i \geq 0) = 1 \quad (a_i \leq 0).$$

We prove this by contradiction. Assume the above function values can be achieved, then $a_i w^\top x_i \leq 0$ for all i. Since there is at least one $a_j > 0$, we know that for this j, $a_j w^\top x_j < 0$. Therefore

$$\sum_{i=1}^{d+1} a_i w^\top x_i < 0,$$

which is a contradiction to (3).

Example of Infinite VC Dimension

Example 15

The binary-valued function class $\mathcal{G} = \{\mathbb{1}(\cos(wz) \geq 0) : w, z \in \mathbb{R}\}$ has infinite VC-dimension.

Given any d, we consider $\left\{z_j = 16^{-j}\pi : j = 1, \ldots, d\right\}$. Let $w = \sum_{j=1}^{d}(1 - b_j)16^j$, with $b_j \in \{0, 1\}$. It is easy to verify that $\mathbb{1}(\cos(w\,z_j) \geq 0) = b_j$. It follows that the set can be shattered by $\mathcal{G}$.

Uniform Convergence with Finite VC-Dimension

Corollary of Theorem 9

Assume $L(\cdot,\cdot) \in \{0,1\}$ is a binary valued loss function. Let

$$\mathcal{G} = \{L(f(w,x),y) : w \in \Omega\},$$

with a finite VC-dimension $\mathrm{VC}(\mathcal{G}) = d$. Then given $\delta \in (0,1)$, with probability at least $1-\delta$ over $\mathcal{S}_n = \{(X_1,Y_1),\ldots,(X_n,Y_n)\}$, the following inequality holds:

$$\begin{aligned}\forall w \in \Omega : & \mathbb{E}_{\mathcal{D}} L(f(w,X),Y) \\ \leq & \frac{1}{n}\sum_{i=1}^{n} L(f(w,X_i),Y_i) + 3\sqrt{\frac{d \ln \max(2, en/d) + \ln(3/\delta)}{2n}}.\end{aligned}$$

This is a direct consequence of Theorem 9 and Sauer's lemma:

$$\ln(N_1(\epsilon = 0, \mathcal{G}, 2n)) \leq d \ln \max(2, en/d)$$

Oracle Inequality Using VC-Dimension

Theorem 16 (Additive Bound in Thm 4.17)

Assume $L(\cdot,\cdot) \in \{0,1\}$ is a binary valued loss function. Let $\mathcal{G} = \{L(f(w,x),y) : w \in \Omega\}$, with a finite VC-dimension $\mathrm{VC}(\mathcal{G}) = d$. Given $n \geq d$, with probability at least $1-\delta$, the ERM solution $\hat{w}$ satisfies:

$$\begin{aligned}\mathbb{E}_{\mathcal{D}} L(f(\hat{w},X),Y) \leq \inf_{w\in\Omega} \mathbb{E}_{\mathcal{D}} L(f(w,X),Y) \\ + \sqrt{\frac{8d\ln(en/d) + 8\ln(4/\delta)}{n}}.\end{aligned}$$

A direct consequence of Corollary 10 and Sauer's lemma.

Multiplicative Oracle Inequality

It is also possible to obtain multiplicative oracle inequality.

Theorem 17 (Multiplicative Bound in Theorem 4.17)

Under the assumptions of Theorem 17. For all $\gamma \in (0,1)$*, with probability at least* $1-\delta$*, the following inequality holds*

$$\begin{aligned}(1-\gamma)^2 \mathbb{E}_{\mathcal{D}} L(f(\hat{w}, X), Y) \leq \inf_{w \in \Omega} & (1+\gamma) \mathbb{E}_{\mathcal{D}} L(f(w, X), Y) \\ & + \frac{(6-3\gamma)(d \ln(en/d) + \ln(4/\delta))}{2\gamma n}.\end{aligned}$$

Margin Bound

VC-dimension of finite dimensional linear classifiers is finite. Theorem 17 can be applied to obtain oracle inequality.

However, for infinite dimensional linear classification problems (e.g. support vector machines), the underlying VC dimension is ∞. In such case, one can try to minimize *margin error* instead of classification error, and obtain generalization error in terms of margin error.

Margin analysis relies on L_∞-covering number analysis at $\epsilon > 0$, which can be finite even for infinite-dimensional classification problems.

Example of Margin Bound

Example 18 (Infinite Dimensional Classification, Expl 4.22)

Consider binary classification with $Y \in \{\pm 1\}$, and linear classifier

$$\left\{ f(w,x) = w^\top \psi(x) : \|w\|_2 \le A \right\},$$

and assume that $\|\psi(x)\|_2 \le B$. Then

$$\mathbb{1}(f(X)Y \le 0) \le \sum_{i=1}^{n} \mathbb{1}(f(w,X_i)Y_i \le \gamma) + O\left(\sqrt{\frac{A^2 B^2 \ln(n + AB/\gamma)}{\gamma^2 n}} \right).$$

One can also obtain the following multiplicative bound for $\gamma \in (0, 0.5]$:

$$\mathbb{1}(f(X)Y \le 0) \le \frac{4}{n} \sum_{i=1}^{n} \mathbb{1}(f(w,X_i)Y_i \le \gamma) + O\left(\frac{A^2 B^2 \ln(n + AB/\gamma)}{\gamma n} \right).$$

Summary (Chapter 4)

- Random partition of data to training versus validation.
- Uniform convergence of training error to validation error using symmetrization.
- Uniform convergence of training error to test error.
- Uniform convergence and oracle inequality using uniform L_1 covering number.
- VC dimension and Sauer's lemma.